Anonymization of data for open science in psychology

Jiří Novák[1,2,3] **Carolin Strobl**[1,2] **Matthias Templ**[2,3]

[1] University of Zürich

[2] University of Applied Sciences and Arts Northwestern Switzerland

[3] Swiss Data Anonymization Competence Center

1. Background

There is a growing demand for more research data to be made openly available. The reproducibility of findings is in crisis, and more openly available data would make research more transparent and accessible.

However, **psychological datasets often include sensitive personal information that necessitates privacy protection**.

OPEN SCIENCE, OPEN ACCESS, OPEN DATA

Research data that results from publicly funded research should be:

- **Findable, Accessible, Interoperable, Reusable** ('**FAIR** principles') [1] [2] therefore replicable, transparent, shareable, trustworthy, verifiable and accountable.
- **As open as possible, as closed as necessary.**

2. Methodology

A key concern with the disclosure of personal data is whether an attacker can gain any new information about an individual.

To enable dissemination and, therefore, to open data, researchers may use methods of **Statistical Disclosure Control (SDC)** [6].

➤ **SDC** is the traditional approach to protect data against re-identification

- **Non-perturbation methods** (partially suppressing or reducing details)
  e.g. Local suppression, Global recoding, Top and bottom coding, Sampling
- **Perturbation methods** (modifying data)
  e.g. Adding Noise, Record swapping, Microaggregaation

➤ **Synthetic data generation** (creating artificial data that mimics the original data and can be safely disseminated)

- **Parametric methods** (statistical models)
- **Non-parametric methods** (machine learning, neural networks - GAN)

3. Example of anonymization

Let's suppose that we are obliged to share data openly while protecting participants' privacy and aligning with open science goals.

➤ **Dataset Description**

- The data for this example is from the Alzheimer's Disease Dataset [4].
- Includes variables such as patient info, demographics, lifestyle, medical history, measurements, symptoms, and diagnosis.

➤ **Anonymization tools**

- Synthetization was performed using the R package synthpop [5].
- For SDC methods we would use package sdcMicro [8] or for complex data package simPop [7].

In our example, we evaluated the utility of synthetic data by comparing original and synthetic datasets on several metrics.

Data utility

Data utility refers to the **usefulness of the data for the intended purpose**. On the other side stands *re-identification risk*, which is the risk that an intruder can link a record in the released data to a specific individual in the population. So, there is a **risk-utility trade-off**. Balancing data utility and privacy is essential. High utility ensures synthetic data's effectiveness for research, while privacy measures minimize re-identification risk.

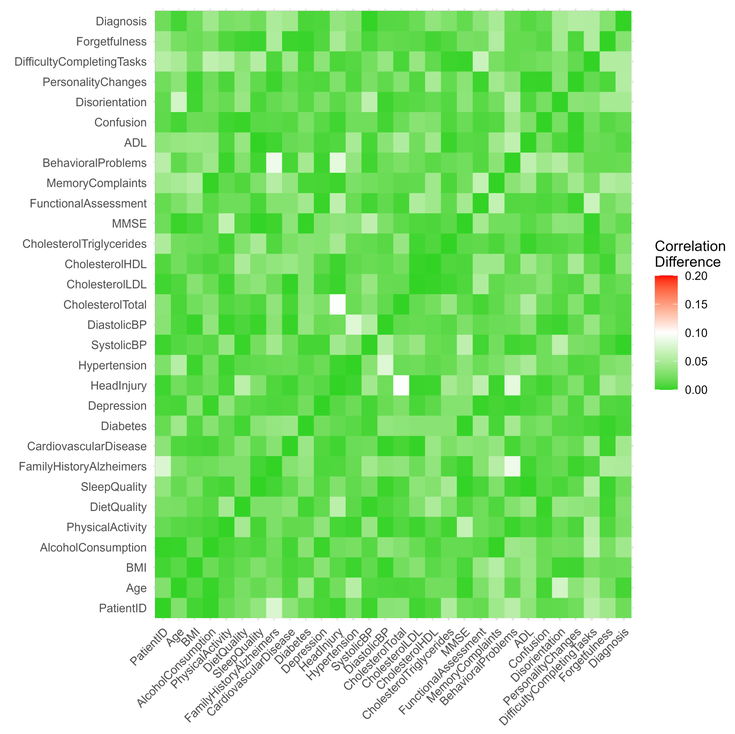

Figure 1: Difference in correlations between Original and Synthetic dataset

The plot illustrates the differences in correlation between the original and synthetic datasets for various numerical variables. Each cell in the heatmap represents the absolute difference in correlation coefficients between the corresponding pairs of variables in the original and synthetic data.

- The predominantly green color indicates that the correlation differences between the original and synthetic datasets are generally small.
- The few white suggest areas where the differences in correlation are slightly higher, but these are sparse and not indicative of significant deviations.

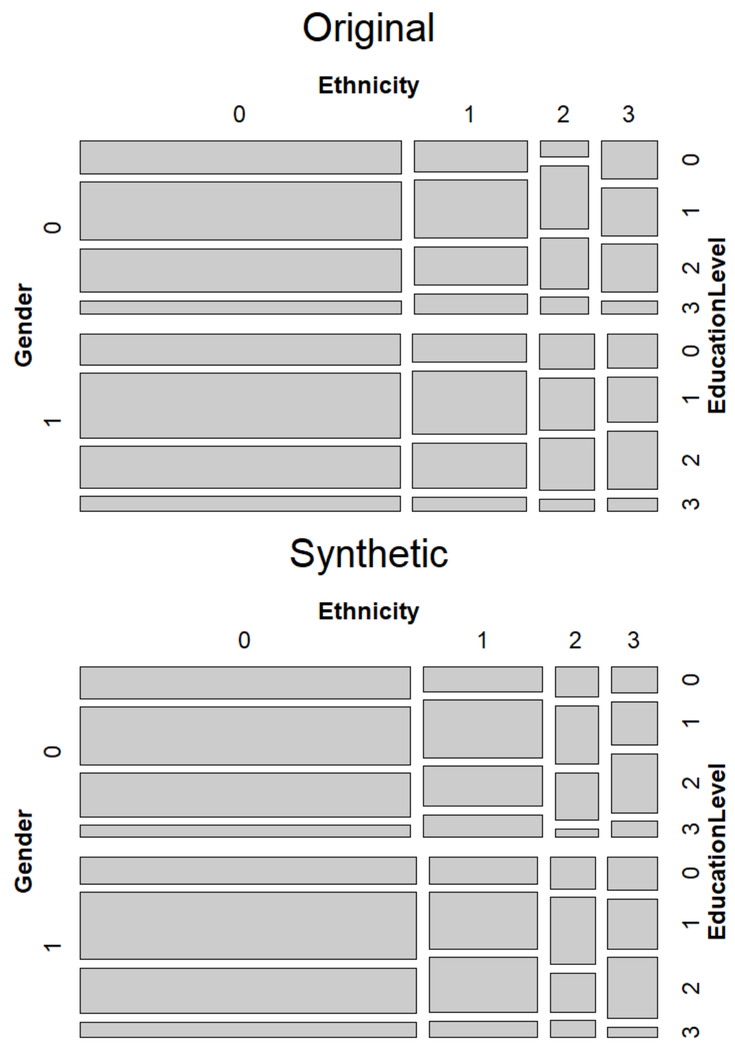

Figure 2: Mosaic plots for selected variables

The mosaic plots display differences in structure for categorical data. In this case, the synthetic and original datasets show highly similar distributions across the variables Gender, Ethnicity, and EducationLevel. This similarity indicates that the synthetic data effectively preserves the relationships and proportions present in the original data, maintaining its analytical utility.

4. Forthcoming Research

The goal of our SNSF*-funded project is developing and implementing innovative tools for generating synthetic longitudinal data with a focus on disclosure risk.

References

[1] European University Association. The European University Association Open Science Agenda 2025, 2022.

[2] European Commission. Commission recommendation (EU) 2018/790 of 25 april 2018 on access to and preservation of scientific information.

[3] A. Hundepool. *Statistical disclosure control*. Wiley series in survey methodology. Wiley, Chichester, West Sussex, United Kingdom, 2012.

[4] Rabie El Kharoua. Alzheimer's disease dataset, 2024.

[5] B. Nowok, G. M Raab, Ch. Dibben, J. Snoke, and C. van Lissa. synthpop: Generating synthetic versions of sensitive microdata for statistical disclosure control.

[6] M. Templ. *Statistical disclosure control for microdata*. Springer Berlin Heidelberg, 2017.

[7] M. Templ, A. Kowarik, B. Meindl, A. Alfons, M. Ribatet, J. Gussenbauer, and S. Fritzmann. simPop: Simulation of complex synthetic data information.

[8] M. Templ, B. Meindl, A. Kowarik, and J. Gussenbauer. sdcMicro: Statistical disclosure control methods for anonymization of data and risk estimation.

*Acknowledgments

This work was funded by the Swiss National Science Foundation (SNSF) with grant number 211751: "*Harnessing event and longitudinal data in industry and health sector through privacy preserving technologies*".